International Journal of Applied Engineering
Research and Development (IJAERD)
ISSN(P): 2250-1584; ISSN(E): 2278-9383
Vol. 4, Issue 2, Apr 2014, 81-88
© TJPRC Pvt. Ltd.

DEFORMED IDENTITY CRIME DETECTION

SOHINI BHATTACHARYA CHAKRABORTY & MOHD ZAFAR SHAIKH

Department of Computer Engineering, C.B.D, Belpada, Navi Mumbai, Maharashtra, India

ABSTRACT

Identity crime is well known, prevalent, and very prominent in our society and credit application fraud is a specific case of identity crime. The existing non-data mining detection systems of business rules and scorecards, and known fraud matching have limitations. To overcome these limitations and combat identity crime in real-time, this paper proposes a new multi-layered detection system complemented with two additional layers: clique Detection and suspicion score Detection. Clique finds real social relationships to reduce the suspicion score, and is tamper-resistant to synthetic social relationships. It is the whitelist-oriented approach on a fixed set of attributes. Suspicion score finds spikes in duplicates to increase the suspicion score, and is probe-resistant for attributes. Research has been carried out on clique and suspicion score with several and huge set of real credit applications. Although this method is specific to credit application fraud detection, but the concept of deformation, together with adaptivity and quality of data discussed in the paper, are general to the design, implementation, and evaluation of all detection systems.

KEYWORDS: Anomaly Detection, Data Mining Based Fraud Detection, Data Stream Mining, Security, Text Attribute

INTRODUCTION

Identity crime is costly, and well known and very prominent in our society. To some extent, synthetic identity fraud refers to the use of plausible but fictitious identities. These are effortless to create but more difficult when it will be applied successfully. Real identity theft refers to illegal use of innocent people's complete identity details. These can be harder to obtain (although large volumes of some identity data might be widely available) but easier to successfully apply. In real scenario identity crime can be treated with a mix of both synthetic and real identity details.

Identity crime is important and vital in now days because there is so much real identity data is available on the Web, and confidential data can be accessed through unsecured mailboxes. It is not difficult for perpetrators to hide their true identities. This incidence can be happened in a field of insurance, credit, and telecommunications fraud, as well as other more serious crimes. Moreover identity crime is prevalent and costly in developed countries that do not have nationally registered identity numbers. Data breaches which involve lost or stolen consumers' identity information can lead to other frauds such as tax returns, home equity, and payment card fraud. As a result, these organizations incur economic damage, such as notification costs, fines, and lost business.

Duplication of Identification

Duplicates (or matches) refer to applications which share common values. There are two types of duplicates: exact (or identical) duplicates have the all same values; near (or approximate) duplicates have some same values (or characters), some similar values with slightly altered spellings, or both or different in their character.

This method argues that each successful occasion fraud pattern is represented by a sudden and sharp spike in duplicates within a short time, relative to the established baseline level. Duplicates are hard to avoid from fraudsters' point of- view because duplicates increases their success rate. So for combating identity fraud in reality this paper proposes a new multilayer detection system complemented with two additional layers clique detection and suspicion score detection. 1st layer finds real social relationship to reduce suspicion score, with fixed set of attribute and suspicion score finds deformation or spikes in duplicates with variable set of attribute.

LITERATURE SURVEY

Many individual data mining algorithm has been designed, implemented and evaluated in fraud detection analysis. There is some pattern in identity crime which can be highly indicative in early symptom in identity fraud especially in synthetic identity crime [3]. In this scheme [14] has ID score risk which gives a combined view of each credit application's characteristics and their similarity to other industry. In another example, it can be detected the application of fraud prevention system [7]. But case based reasoning (CBR) is the only known prior publication in the screening of credit application [8]. My proposed approach which monitors the significant increase or decrease in amount of something important is similar in concept to credit transactional fraud detection and bio terrorism detection. In case of fraud detection peer group analysis [2] monitors inter account behavior over time. It compares the cumulative mean weekly amount between a target account and other similar accounts (peer group) at subsequent time points. Bayesian Network [4] uncovers simulated anthrax attack from real emergency department data. Surveys algorithms [5] are used for finding suspicious activity in time for disease outbreaks. [9] Uses time series analysis to track early symptoms of synthetic anthrax outbreaks from daily sales of retails medication .Control chart based statistics, exponential weighted moving averages and generalized linear models were tested on the same bio terrorism detection of data and alert rate [15]. In addition my proposed algorithm suspicion score detection is similar to change point detection in bio surveillance research, which maintains the cumulative sum (CUSUM) of positive derivation from the mean [13].

Limitation of Existing Approach

In the real-time credit application fraud detection domain, this paper argues against the use of classification (or supervised) algorithms which use class labels. In addition to the problems of using known frauds, these algorithms, such as logistic regression, neural networks, or Support Vector Machines (SVM), cannot achieve scalability or handle the extreme imbalanced class [11] in credit application data streams. As fraud and legal behavior changes frequently, the classifiers will deteriorate rapidly and the supervised classification algorithms will need to be trained on the new data. But the training time is too long for real-time credit application fraud detection because the new training data has too many derived numerical attributes (converted from the original, sparse string attributes) and too few known frauds.

Proposed System

The new methods are based on white-listing and detecting suspicion score of similar applications. White-listing or clique detection uses real social relationships on a fixed set of attributes. This reduces false positives by lowering some suspicion scores. Detecting spikes in duplicates on a variable set of attributes, increases true positives by adjusting suspicion scores appropriately. Throughout this paper, data mining is defined as the real-time search for patterns in a principled (or systematic) fashion. These patterns can be highly indicative of early symptoms in identity crime, especially synthetic identity fraud [10].

Objective of Proposed Method

The main objective of this method is to achieve deformation or resilience by adding two new, real-time, data mining-based layers to complement the two existing non-data mining layers. These new layers will improve detection of fraudulent applications because the detection system can detect more types of attacks, better account for changing legal behavior, and remove the redundant attributes. These new layers are not human resource intensive. They represent patterns in a score where the higher the score for an application, the higher the suspicion of fraud (or anomaly). In this way, only the highest scores require human intervention. Crucially, these two layers are unsupervised algorithms which are not completely dependent on known frauds but use them only for evaluation. The first new layer is clique detection: the white list-oriented approach on a fixed set of attributes. To complement and strengthen CLIQUE, the second new layer is Suspicion score Detection (SD): the attribute-oriented approach on a variable-size set of attributes. The second contribution is the significant extension of knowledge in credit application fraud detection because publications in this area are rare. In addition, this research uses the key ideas from other related domains to design the credit application fraud detection algorithms.

Finally, at the end of this technique is the recommendation of credit application fraud detection as one of the many solutions to identity crime. Being at the first stage of the credit life cycle, credit application fraud detection also prevents some credit transactional fraud.

METHODOLOGY

This section is divided into four subsections to systematically explain the clique detection algorithm (first two subsections) and the suspicion score detection algorithm (last two subsections). Each subsection commences with a clearer discussion about its purposes.

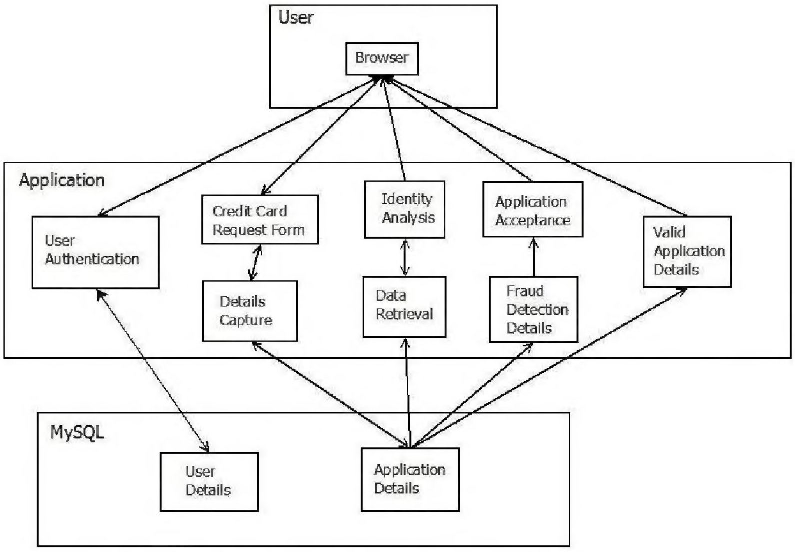

Figure 1: System Architecture of Proposed System

Basic Steps of this Mechanism

- Clique detection method calculates the suspicion score for each attribute for the current application with respect to the existing applications in terms of their similarities.

- This suspicion score is compared with valid relationships defined in the system and utilized to decrease it to identify the exact entries suitable for identity detection.
- Calculate score of every single link with previous applications are performed.
- Calculate multiple link score based on initial score and previous applications score.
- Based on above values it will generate the updated valid set of values for relationships.
- Suspicion score detects the deviation in values above a threshold value by dividing the applications into sets of data.
- The deviation score is identified here based on each attribute here.
- The attributes that is required for Suspicion score is identified.
- Based on all selected attributes the overall suspicion score is calculated.
- The weights associated with each of the attributes is calculated and utilized in the CLIQUE execution.

Inputs

vi (current application)

W number of vj (moving window)

_a, link-type (link-types in current white list)

S_ similarity (string similarity threshold)

S attribute (attribute threshold)

η (exact duplicate filter)

α (exponential smoothing factor)

S input (input size threshold)

Outputs

Sc (vi) (suspicion score)

Same or new parameter value

New whit list

Clique Detection Basic Algorithm Layout

Step 1: Multi-attribute link establishment [match vi against W number of vj to determine if a single attribute exceeds S _ similarity; and create multi-attribute links if near duplicates' similarity exceeds T attribute or an exact duplicates' time difference exceeds η]

Step 2: Single-link score value [calculate single-link score by matching Step 1's multi-attribute links against _a, link-type]

Step 3: Single-link average previous score formation [calculate average previous scores from Step 1's linked previous applications]

Step 4: Multiple-links score formation [calculate Sc(vi) based on weighted average (using α) of Step 2's link scores and Step 3's average previous scores]

Step 5: White list change [determine new white list at end of the result].

Multi Attribute Link Count by Jaro-Winkler Formula

The Jaro–Winkler distance metric is designed and best suited for short strings such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match. The Jaro distance of two given strings and is the Jaro distance d_j of two given strings $s_1$and $s_2$is

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

Where:

$m$is the number of matching characters

Two characters from $s_1$and $s_2$respectively, are considered matching only if they are the same and not farther than $\left\lfloor \frac{\max(|s_1|,|s_2|)}{2} \right\rfloor - 1$.

Suspicion Score Detection Method

Inputs

vi (current application)

W number of vj (moving window)

t (current step)

S similarity (string similarity threshold)

θ (time difference filter)

α (exponential smoothing factor)

Outputs

S (vi) (suspicion score)

wk (attribute weight)

Suspicion Score Detection Algorithm

Step 1: Single-step scaled counts measurement [match vi against W number of vj to determine if a single value exceeds S similarity and its time difference exceeds θ]

Step 2: Single-value spike or deformation detection [calculate current value's score based on weighted average (using α) of t Step 1's scaled matches]

Step 3: Multiple-values score [calculate S (vi) from Step 2's value scores and Step 4's wk]

Step 4: Suspicion score attributes selection [determine wk for Spike at end of gx]

Step 5: CLIQUE attributes weights change [determine wk for CLIQUE at end of gx]

Module Based Application

This application has been experimented on credit based application but in case of all detection system it can be implemented and applied. In this method GUI module is designed for this module facilitates authentication of various users and thereby providing access to selected users within the system

Apply for Credit Card

This feature will allow various users to apply for credit cards using various details required in the application to perform invalid application scenario.

Track Details

The details of all applications is tracked and utilized in detection mechanism.

Change Password

This module will facilitate changing the password details of the user.

Application Validation

The applications will be analyzed using Fraud Detection Techniques to identify the identity conflict scenarios with the system and displaying it to the admin.

Application Acceptance

This will be a admin module displaying the conflicts in the system and finally allowing admin to reject or accept these applications.

Following will be variations (queries) that the algorithm suggests to detect:

- Two applications consisting of similar data but different First and Last Name.
- Two applications consisting of similar data but different City only.
- Two applications consisting of similar data but different Secondary Contact Number.
- Two applications consisting of similar data but different Primary Contact Number.

Analysis and Discussion of this Method

This experiment is carried out on Windows XP operating system on java platform. Here experiment is carried out on Tomcat application server5.0.Here server side script is Java server page and scripts are java script. All data are stored in My Sql database. The front end application is HTML, Java, Jsp, XML and database connectivity is JDBC. Identity data - Real Application Dataset (RADS)Substantial identity crime can be found in private and commercial databases containing information collected about customers, employees, suppliers, and rule violators. The same situation occurs in public and government-regulated databases such as birth, death, patient and disease registries; taxpayers, residents' address, bankruptcy, and criminals lists. To reduce identity crime, the most important textual identity attributes such as personal name, (first name, last name) address (Street name ,city, pin code) State name, and personal uniqueness like primary phone no and secondary phone no are important. The most important identity attributes differ from database to database. They are least likely to be manipulated, and are easiest to collect and investigate. They also have the least missing values, least spelling and transcription errors, and have no encrypted values Daily application volume for population of dataset for the time being in present and in future.

Advantages of Proposed System

Much work in credit application fraud detection remains proprietary and exact performance figures unpublished, therefore there is no way to compare the clique and SD algorithms against their leading industry methods and techniques. For example, one existing mechanism has ID Score-Risk which gives a combined view of each credit application's characteristics and their similarity to other industry provided for analyzing Web identity's characteristics. But this mechanism does not facilitate the system to accurately detect the identity deviation provided by the Suspicious score Detection mechanism and therefore lower accuracy in detection.

CONCLUSIONS

The main focus of this paper is deformation and spike of deformation find out Identity Crime Detection; in other words, the real-time search for patterns in a multi-layered and principled fashion, to safeguard credit applications at the first stage of the credit life cycle. This paper describes an important domain that has many problems relevant to other data mining research. It has documented the development and evaluation in the data mining layers of defense for a real-time credit application fraud detection system. In doing so, this research produced three concepts which increase the detection system's effectiveness (at the expense of some efficiency). These concepts are resilience (multi-layer defense), adaptivity (accounts for changing fraud and legal behavior), and quality data (real-time removal of data errors). These concepts are fundamental to the design, implementation, and evaluation of all fraud detection, adversarial-related detection, and identity crime-related detection systems. The implementation of CLIQUE and suspicion score algorithms is practical because these algorithms are designed for actual use to complement the existing detection system.

REFERENCES

1. Bifet, A. and Kirk by, R. 2009. Massive Online Analysis, Technical Manual, University of Waikato.

2. Bolton, R. and Hand, D. 2001. Unsupervised Profiling Methods for Fraud Detection, Proc. of CSCC01.

3. Oscherwitz, T. 2005. Synthetic Identity Fraud: Unseen Identity Challenge, Bank Security News 3: p.7.

4. Wong, W., Moore, A., Cooper, G. and Wagner, M. 2003. Bayesian Network Anomaly Pattern Detection for Detecting Disease Outbreaks, Proc. of ICML03. ISBN: 1-57735-189-4.

5. Wong, W. 2004. Data Mining for Early Disease Outbreak Detection, PhD thesis, Carnegie Mellon University.

6. Cortes, C., Pregibon, D. and Volinsky, C. 2003. Computational methods for dynamic graphs, Journal of Computational and Graphical Statistics 12(4): pp. 950-970. DOI: 10.1198/ 1061860032742.

7. Experian. 2008. Experian Detect: Application Fraud Prevention System. Whitepaper, http://www.experian.com/products/pdf/experian detect.pdf.

8. Wheeler, R. and Aitken, S. 2000. Multiple Algorithms for Fraud Detection, Knowledge-Based Systems 13(3): pp. 93-99. DOI: 10.1016/S0950-7051(00)00050-2.

9. Goldenberg, A., Shmueli, G. and Caruana, R. 2002. Using Grocery Sales Data for the Detection of Bio-Terrorist Attacks, Statistical Medicine.

10. Gordon, G., Rebovich, D, Choo, K. and Gordon, J. 2007. Identity Fraud Trends and Patterns: Building a Data-Based Foundation for Proactive Enforcement, Center for Identity Management and Information Protection, Utica College.

11. Hand, D. 2006. Classifier Technology and the Illusion of Progress, Statistical Science 21(1): pp. 1-15. DOI: 10.1214/088342306000000060.

12. Head, B. 2006. Biometrics Gets in the Picture, Information Age August-September: pp. 10-11.

13. Hutwagner, L., Thompson, W, Seeman, G., Treadwell, T. 2006. The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS), Journal of Urban Health 80: pp. 89-96.PMID: 12791783.

14. ID Analytics. 2008. ID Score-Risk: Gain Greater Visibility into Individual Identity Risk. Unpublished.

15. Jackson, M., Baer, A., Painter, I. and Duchin, J. 2007. A Simulation Study Comparing Aberration Detection Algorithms for Syndrome Surveillance, BMC Medical Informatics and Decision Making 7(6). DOI: 10.1186/1472-6947-7-6.